ENT-ICIPATE

Checkpoint #2

Team 7: Alessandro Carrabs, Xiao Quan Ji, Samesun Singh

TABLE OF CONTENTS

OBJECTIVES

POST-OPERATIVE COMPLICATIONS FOR ENT PATIENTS SUCH AS NOSOCOMIAL INFECTION AND PHARYNGO-/ORO-CUTANUOUS FISTULA ARE RARE BUT HAVE STRONG IMPACT ON PATIENT'S SAFETY AND HOSPITAL RESOURCES

WE WANT TO:

- BUILD A ML MODEL TO ESTIMATE EACH PATIENT'S RISK OF SOME COMPLICATIONS
- COMPARE DIFFERENT MODELS AND STRATEGY TO IDENTIFY A ROBUST AND INTERPRETABLE SOLUTION

VALUE PROPOSITION

Transforms raw clinical data from 550+ ENT oncology patients into actionable insights, enabling earlier identification of high-risk cases and improving post-surgical safety.

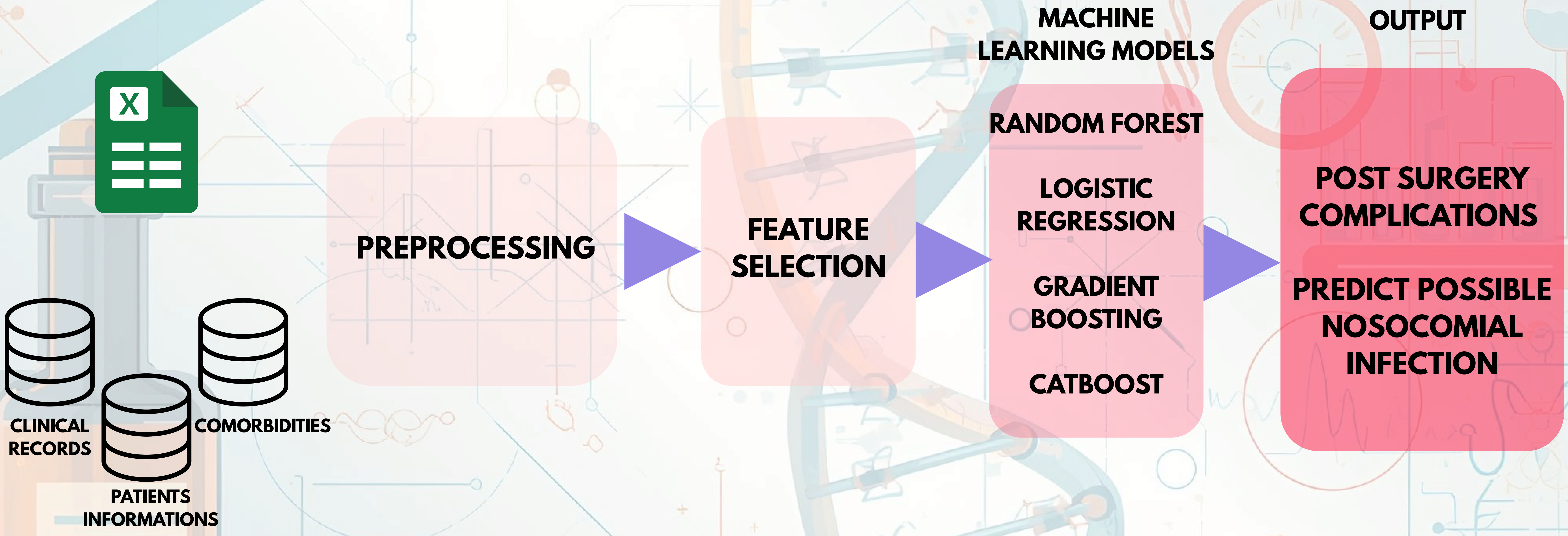
Builds a data-driven infrastructure that standardizes heterogeneous medical records, integrates them into a predictive pipeline, and lays the foundation for scalable AI-assisted clinical workflows.

RESEARCH QUESTIONS

CAN MACHINE LEARNING MODELS ACCURATELY PREDICT POST-SURGICAL COMPLICATIONS IN ENT ONCOLOGY PATIENTS, EVEN WHEN SUCH EVENTS ARE RARE?

WHICH CLINICAL AND SURGICAL FEATURES CONTRIBUTE MOST TO THE RISK OF POST-OPERATIVE COMPLICATIONS?

FUNCTIONAL DIAGRAM



WHAT WE ARE DOING

STEP 1 → DATA UNDERSTANDING AND CLEANING

IDENTIFY INCONSISTENT FORMATS
DROPPING IRRELEVANT COLUMNS
FIXING INCOSTINTENT LABELS

STEP 2 → HANDLING DATA QUALITY

DETECT MISSING VALUES
IMPUTE NUMERICAL AND CATEGORICAL FIELDS

STEP 3 → MODEL PREPARATION

STRATIFIED TRAIN-TEST SPLIT
IMBALANCE HANDLING
READY-TO-TRAIN FEATURE MATRIX

DATA EXPLORATION

COMPOSITION OF THE DATASET

574 Patients | **64 Features**

- Patients demographics and habits
- Comorbidities
- Surgical and operative treatments
- Targets (binary)
  Fistula
  Nosocomial infection

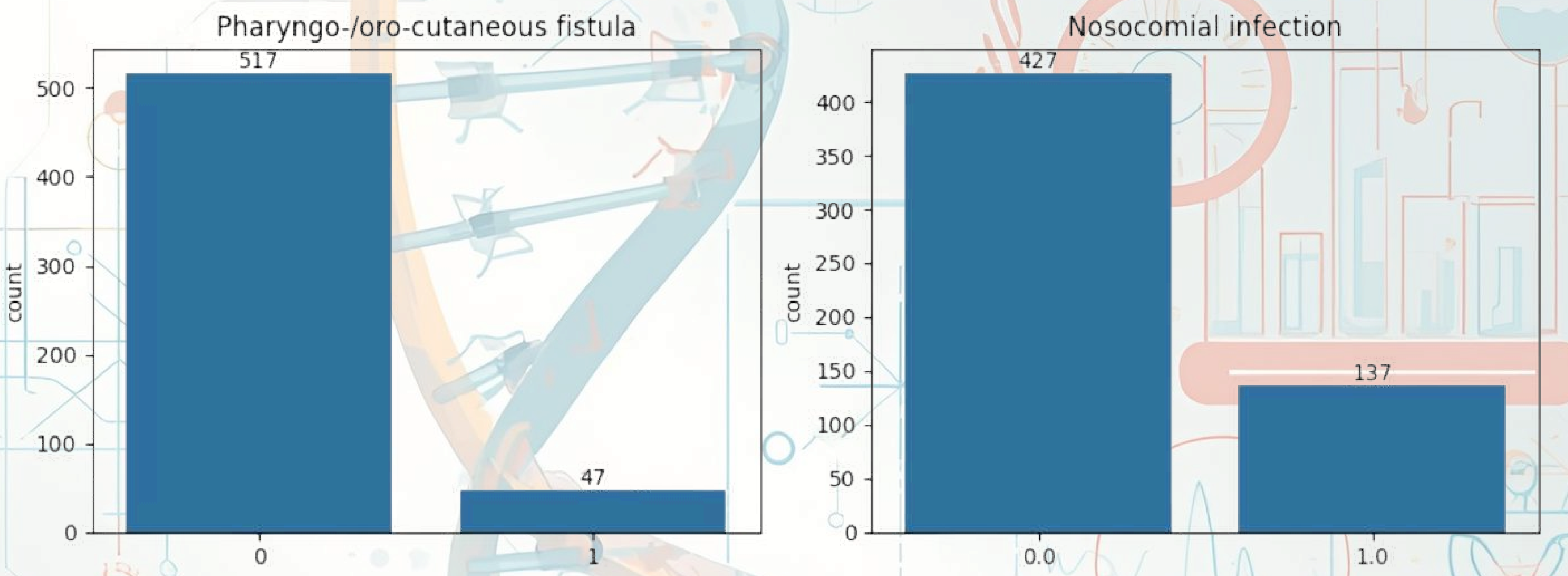
Clinical Data Challenges

- High heterogeneity across variables
- Many categorical features stored as free text
- TNM staging unstructured and inconsistently encoded
- Hidden missing values not immediately detectable

IMBALANCE ASSESTMENT

The dataset shows a **strong class imbalance** in both target variables

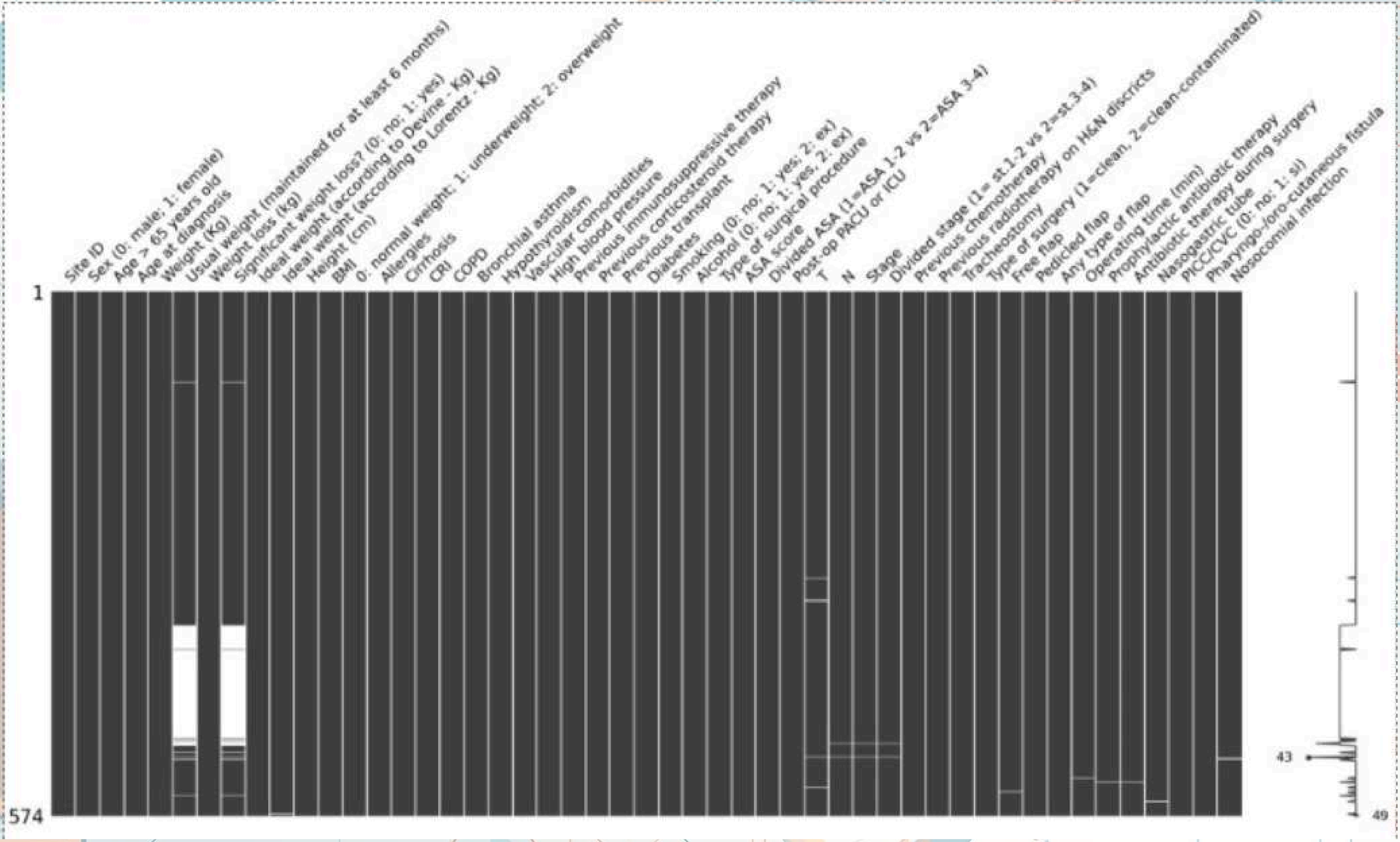
These postoperative complications are rare events, which is expected from a **real clinical setting**.



IMBALANCED DATA CAN LEAD MACHINE-LEARNING MODELS TO FAVOR THE MAJORITY CLASS, REDUCING THE ABILITY TO DETECT HIGH-RISK PATIENTS.

DATA PREPROCESSING

MISSING VALUES



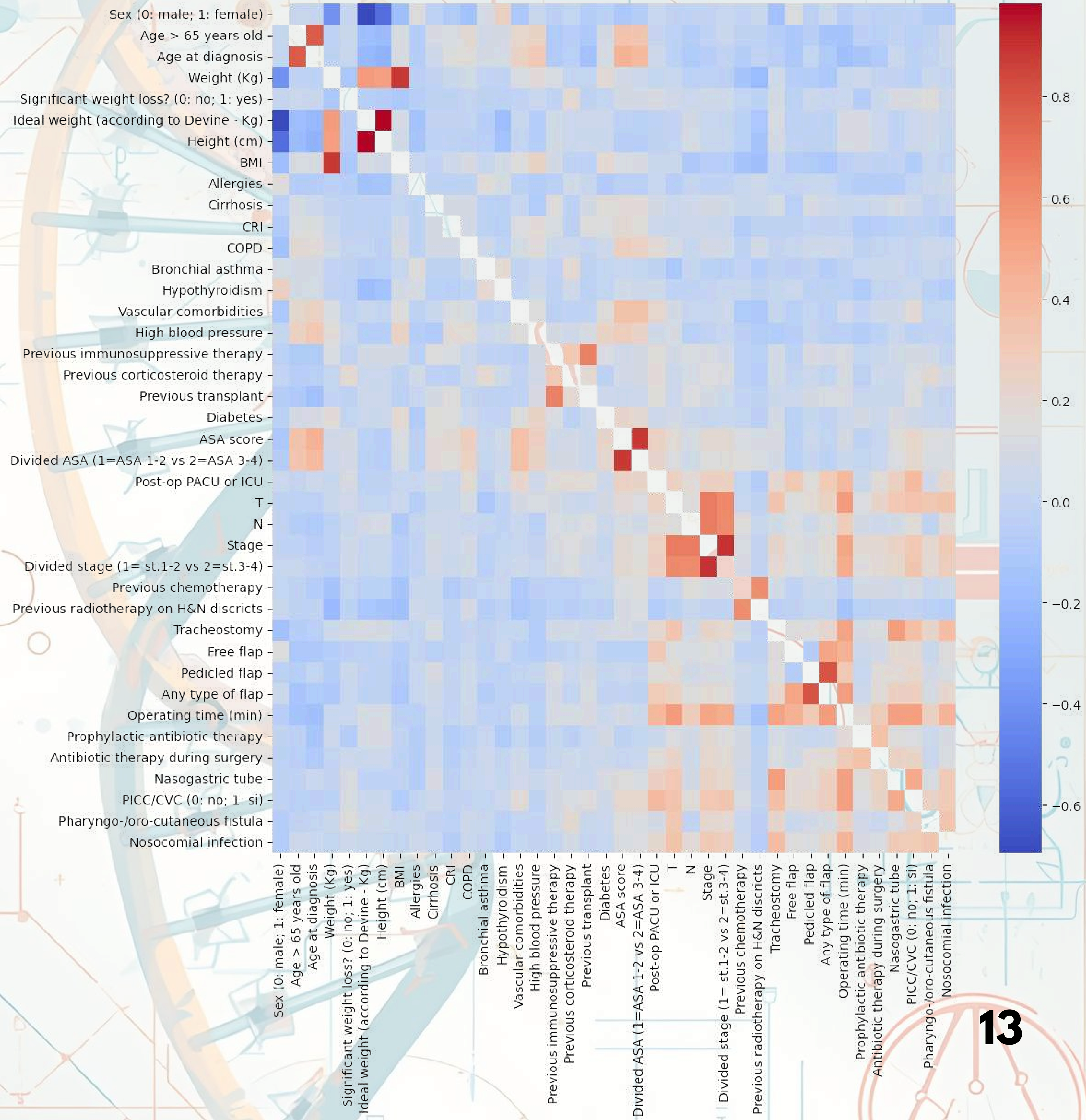
CORRELATION MATRIX

Strong pairwise correlations.
This reduces overall interpretability.

Some groups of variables represent the same clinical concept encoded in multiple ways. This provides redundant information to the model, which does not improve predictive power and can even reduce model stability.

Affected variables:

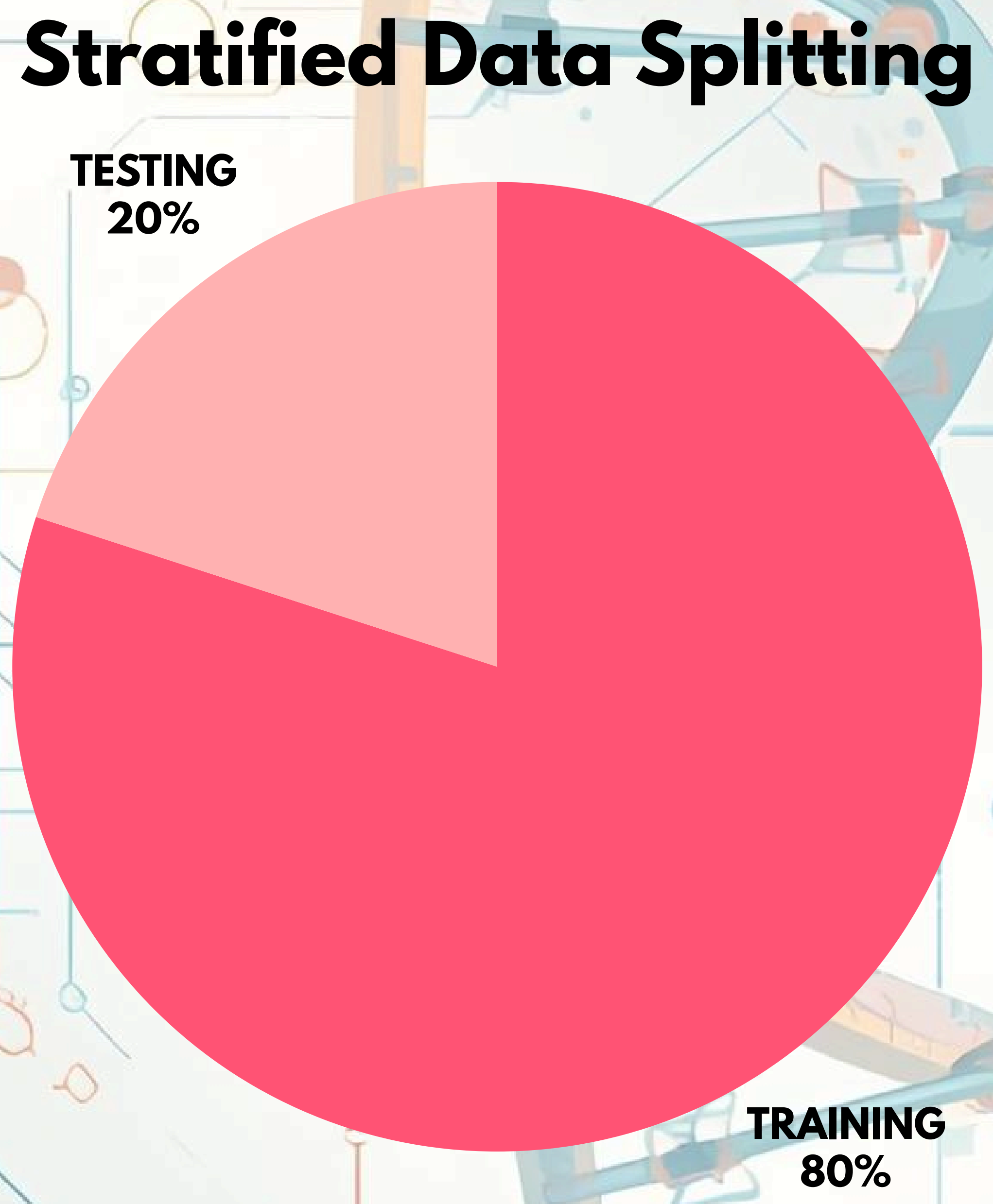
- **Weight informations**
- **ASA and Stage score encoded in two ways.**
- **Age information**
- **Presence of flap**



FEATURE - TARGET CORRELATION

IMPLEMENTATION

DATA SPLIT



STRATIFIED: KEEPS THE **SAME CLASS DISTRIBUTION** IN BOTH TRAIN AND TEST SETS

RECALL: PRIMARY METRIC

HOW MANY HIGH-RISK PATIENTS WE ACTUALLY CATCH?

- **Missing a true complication (false negative) is more dangerous than raising an unnecessary alert**
- **Accuracy is important, but in this case misleading: a model that predicts “no complication” would have very high accuracy due to the dataset, but NO clinical value**

BEST RESULTS

EVALUATION METRIC	RANDOM FOREST for TARGET: Pharyngo-/oro-cutanoeous fistula	CATBOOST for TARGET: Nosocomial infection
ROC-AUC	0.913	0.813
F1	0.516	0.625
Recall	0.889	0.741
Precision	0.364	0.541

For class "1"

NEXT STEPS

MODEL COMPARISON

Test further models and techniques, and compare them

CATEGORICAL ENCODING & FEATURE ENGINEERING

Explore additional feature engineering techniques and try alternative encodings for categorical features to evaluate whether performance improves

INTERPRETABILITY

Analyse feature importance of the best model to provide interpretable insights

THANK YOU

FOR THE ATTENTION

QUESTIONS?